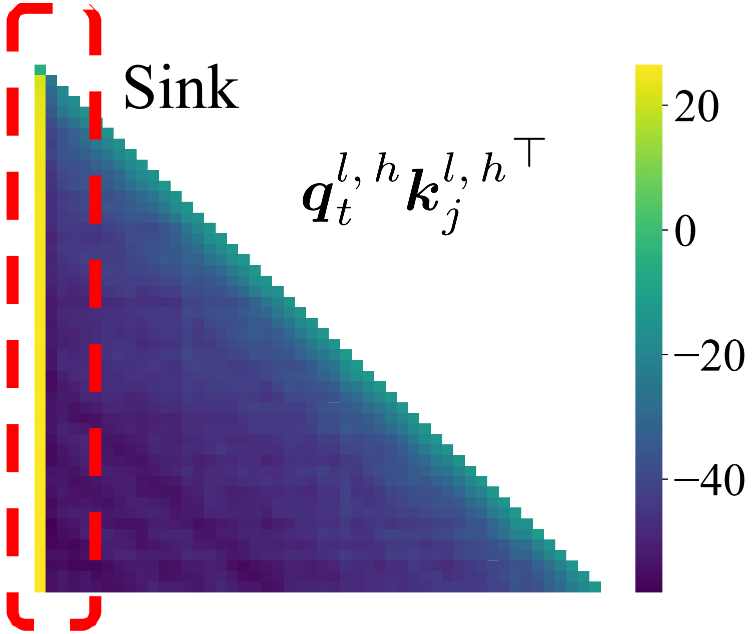
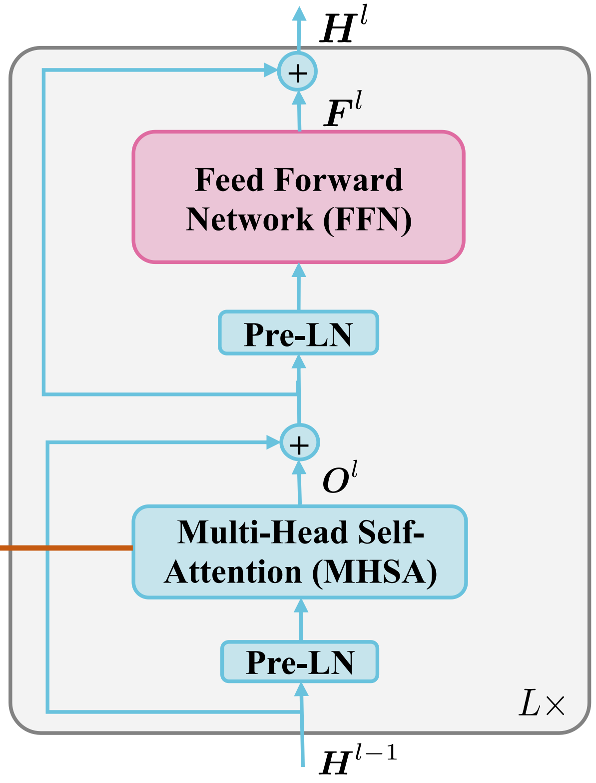
When Attention Sink Emerges in Language Models: An Empirical View

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu,
Fengzhuo Zhang, Cunxiao Du, Ye Wang, Min Lin

What is attention sink?

- Attention sink refers to that Language Models (LMs) assign significant attention to the first token (Xiao et al. 2024)

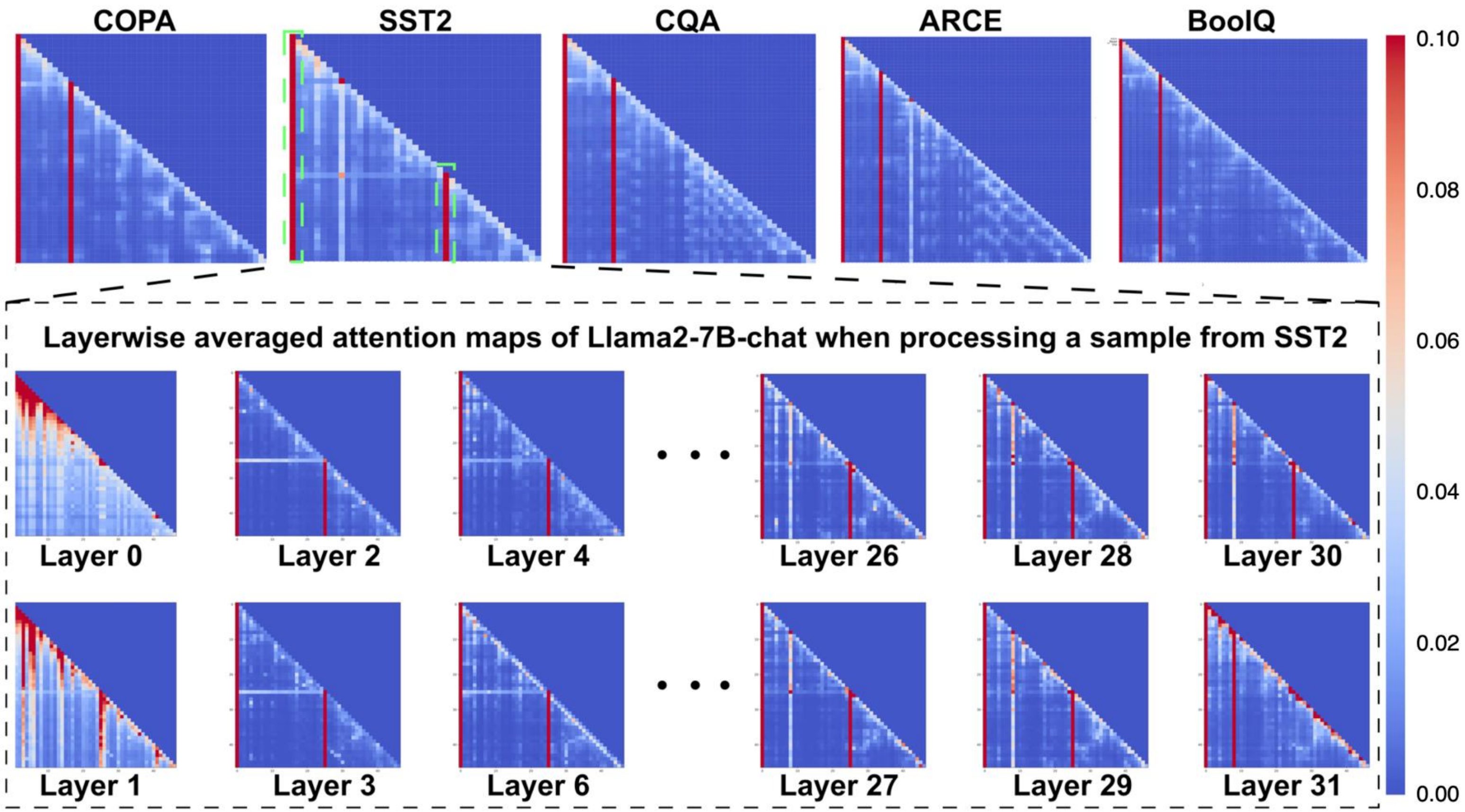




Xiao et al. Efficient Streaming Language Models with Attention Sinks. ICLR 2024

What is attention sink?

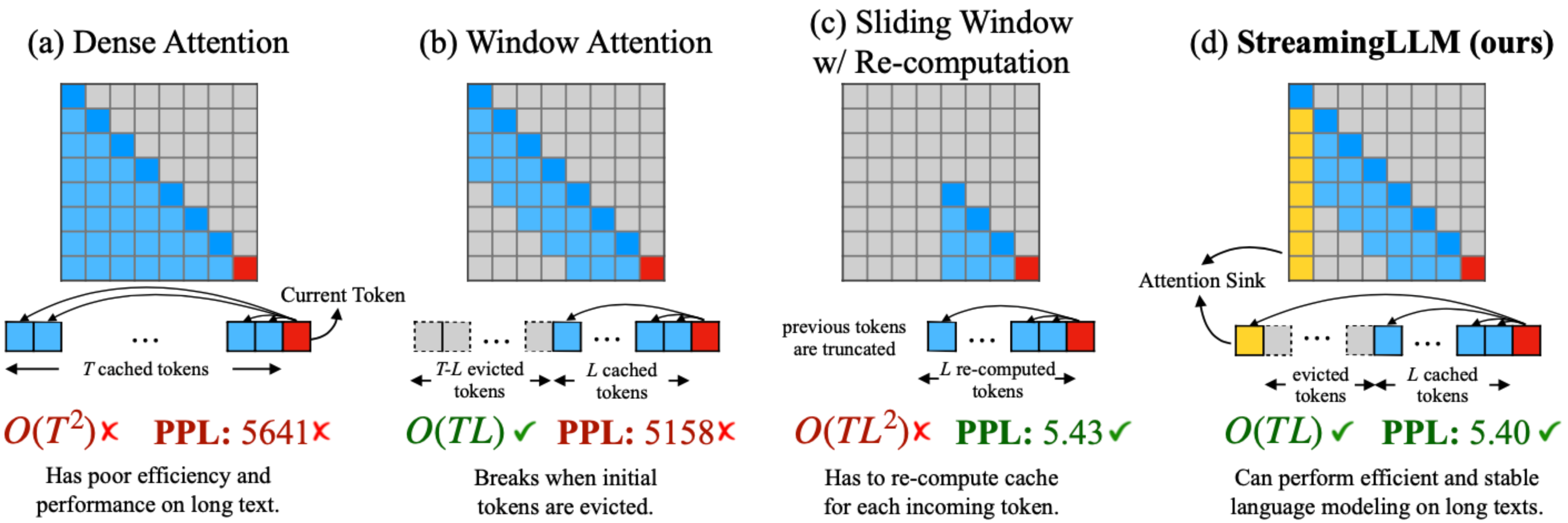
- In some cases, specific tokens may become sink tokens (Yu et al. 2024)



Yu et al. Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration. ICML 2024

What can we do with attention sink?

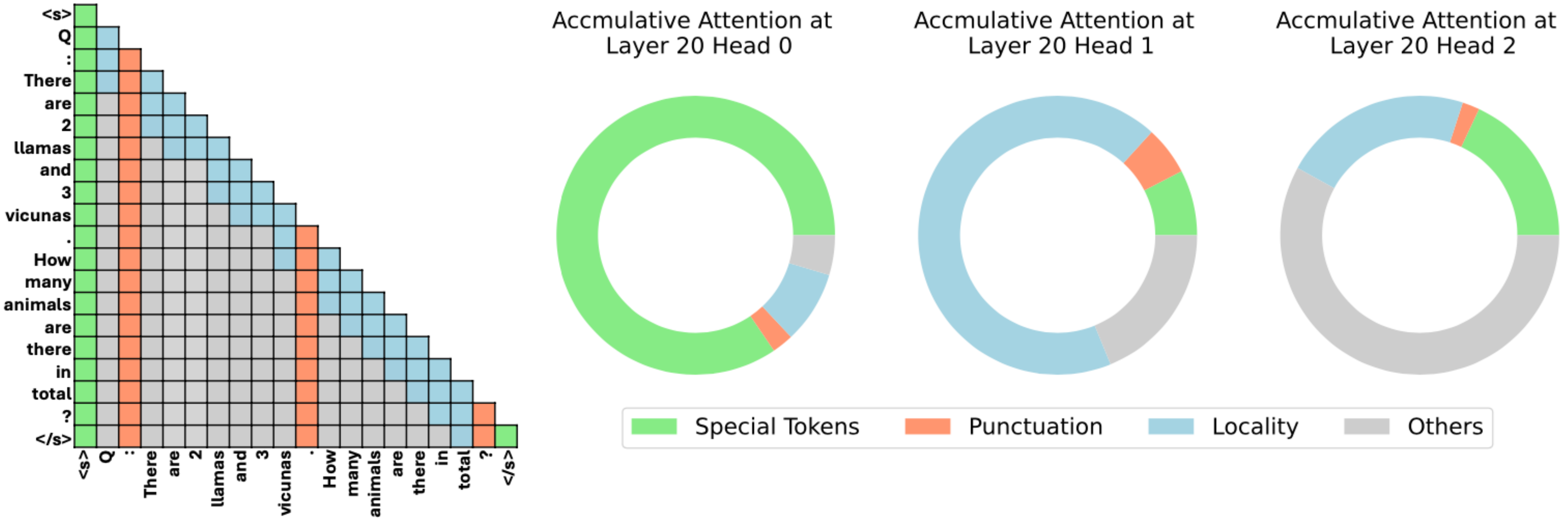
- Long context understanding / generation by only computing the attention on the sink token and recent tokens (Xiao et al. 2024)



Xiao et al. Efficient Streaming Language Models with Attention Sinks. ICLR 2024

What can we do with attention sink?

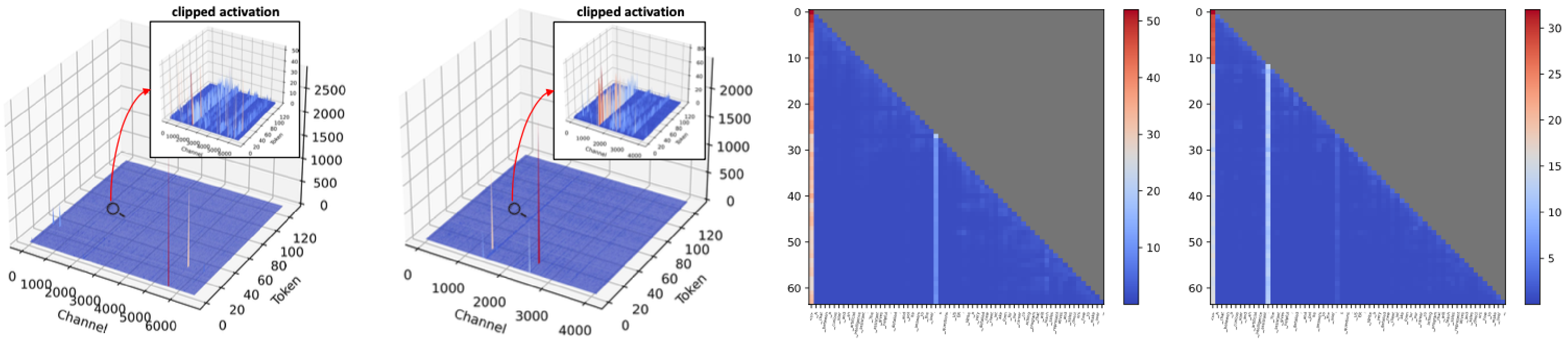
- KV cache compression by only constructing the KV cache of special tokens (including sink tokens) and recent tokens (Ge et al. 2024)



Ge et al. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. ICLR 2024

What can we do with attention sink?

- Model quantization by preserving the KV cache of sink tokens with full precision (Liu et al. 2024)



(a) Output activations of LLaMA-30B Layer 24

(b) Output activations of LLaMA-2-7B Layer 24

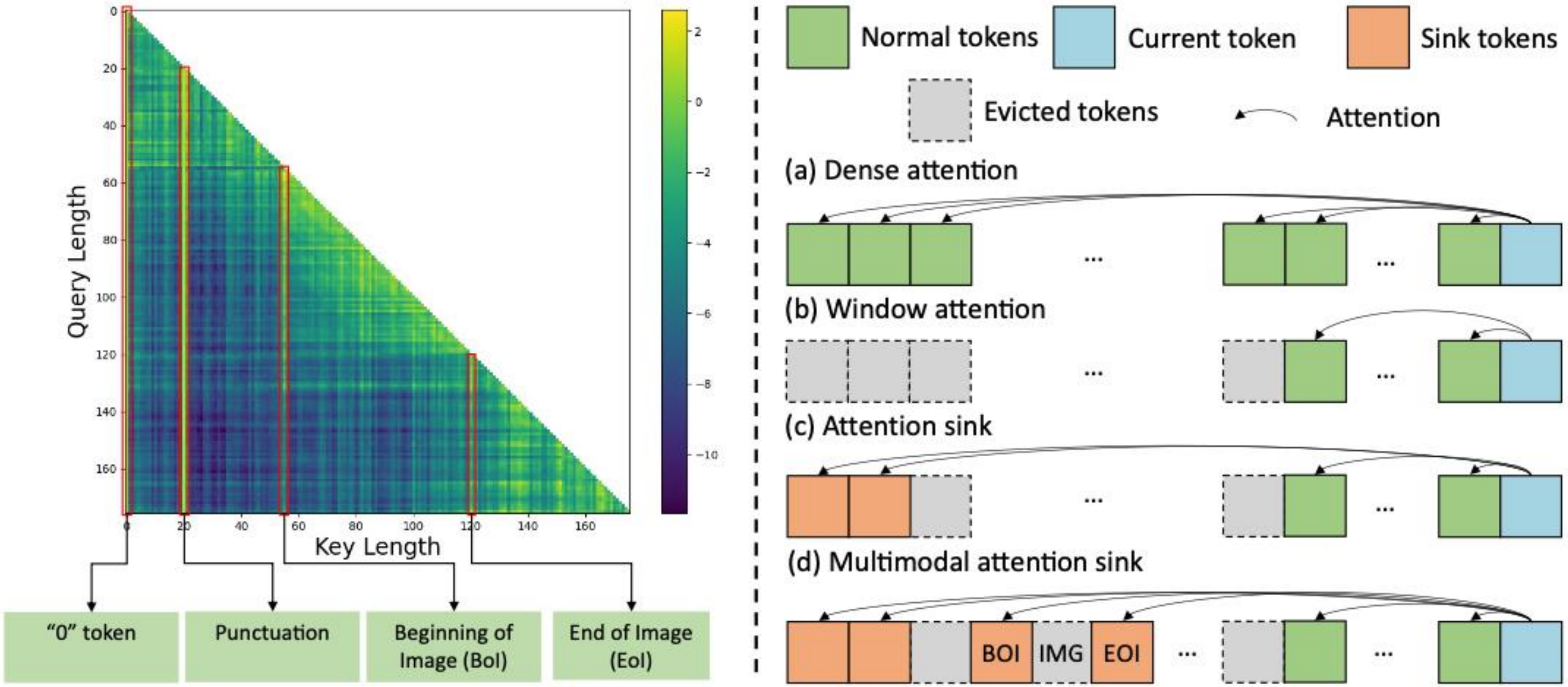
(c) Attention map of LLaMA-30B Layer 24

(d) Attention map of LLaMA-2-7B Layer 24

Liu et al. IntactKV: Improving Large Language Model Quantization by Keeping Pivot Tokens Intact. ACL Findings 2024

What can we do with attention sink?

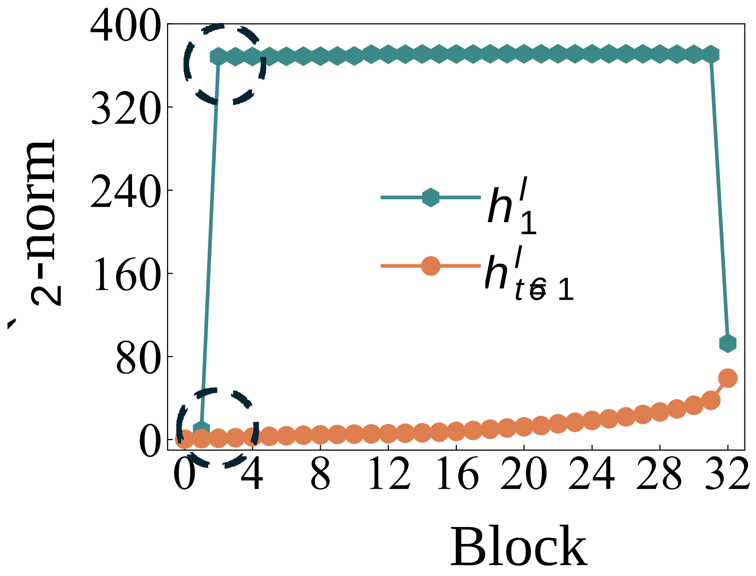
- Multi-model language modeling by considering attention sink (Yang et al. 2024

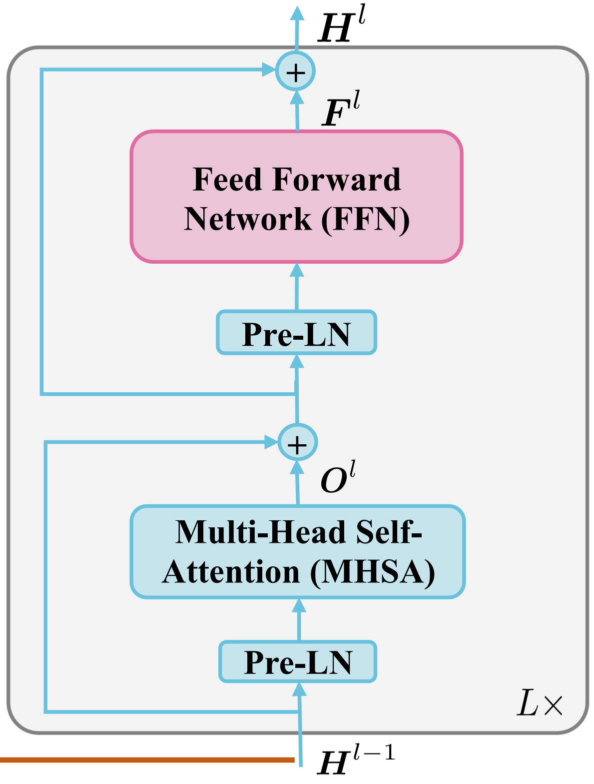


Yang et al. SEED-Story: Multimodal Long Story Generation with Large Language Model. Arxiv 2024

Mechanism of attention sink

- Massive Activations in hidden states of sink token: its L2-norm is significantly larger than that of other tokens (Cancedda 2024; Sun et al. 2024)





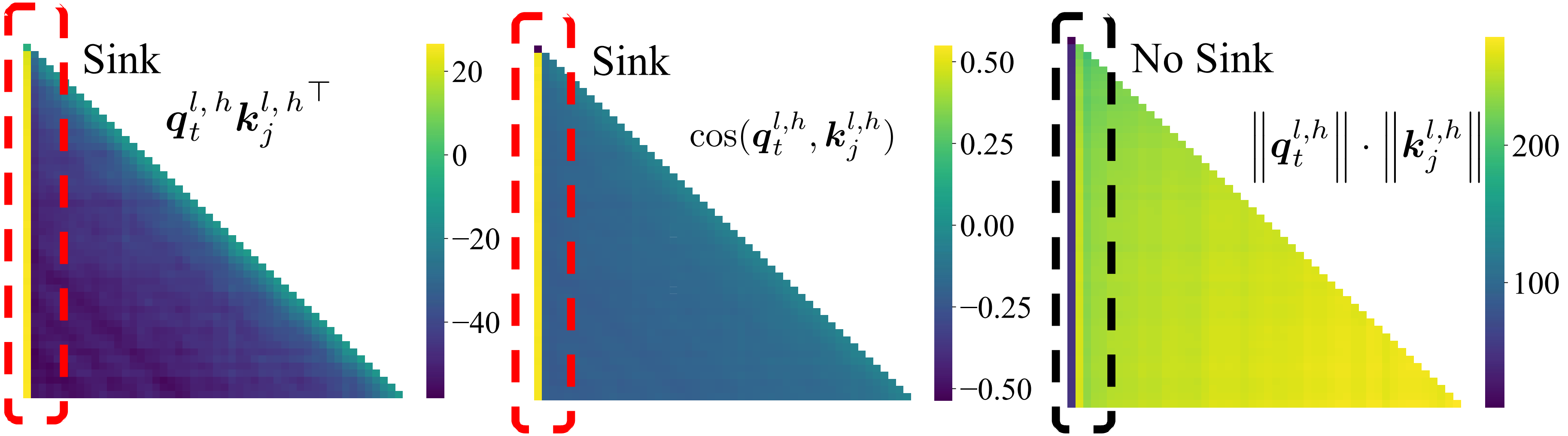
Cancedda, Nicola. Spectral filters, dark signals, and attention sinks. ACL 2024
Sun et al. Massive activations in large language models. COLM 2024

Mechanism of attention sink

- We find that QK angle matters for attention sink

Attention sink $\quad \boldsymbol{q}_t^{l,h} \boldsymbol{k}_1^{l,h\top} \gg \boldsymbol{q}_t^{l,h} \boldsymbol{k}_{j\neq 1}^{l,h\top}$

QK angle $\quad \cos(\boldsymbol{q}_t^{l,h}, \boldsymbol{k}_1^{l,h}) \gg \cos(\boldsymbol{q}_t^{l,h}, \boldsymbol{k}_{j\neq 1}^{l,h})$

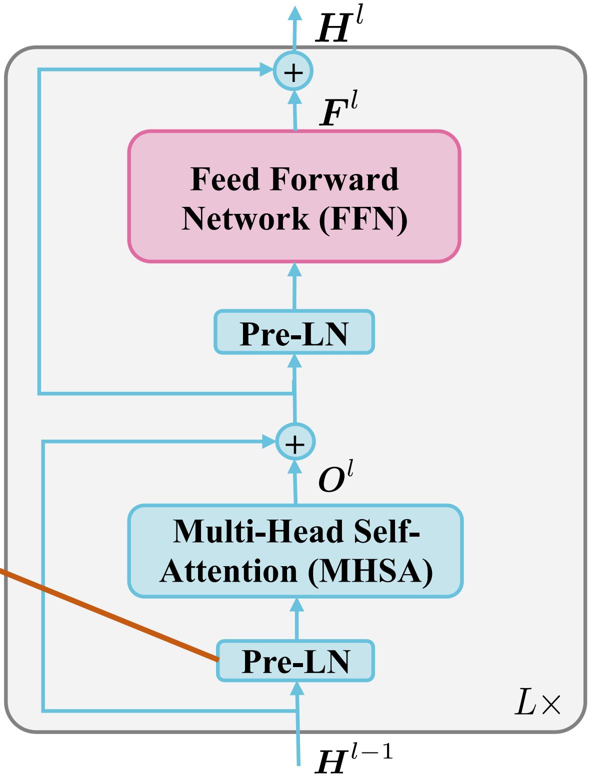


Key of the sink token is distributed in a different manifold

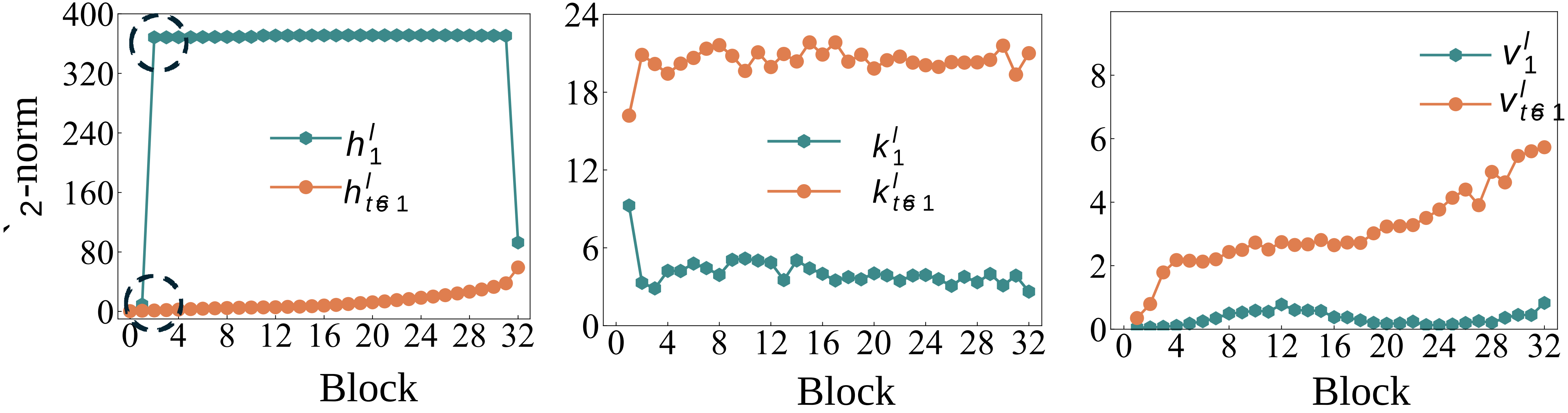
Mechanism of attention sink

- Why massive activations?

Layer norm retains values for specific dimensions for key of sink token



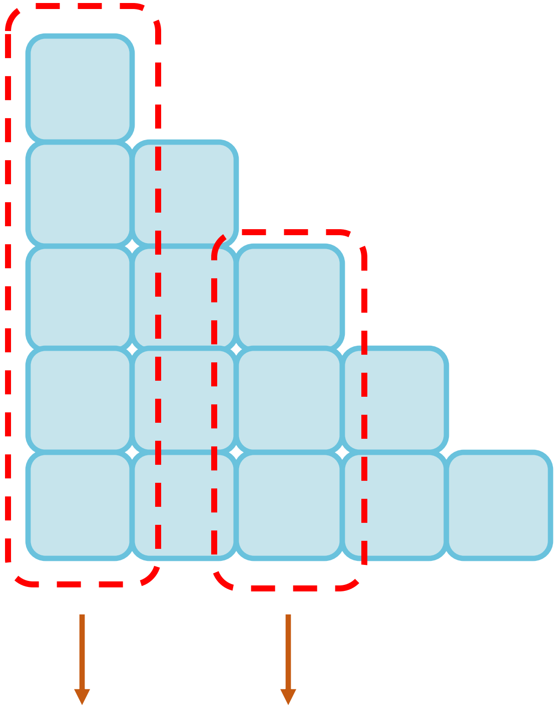
- Special property of KV of sink token



How to measure attention sink?

- Attention scores of the first token are significantly larger than others

Average



Importance score $\alpha_k^{l,h}$

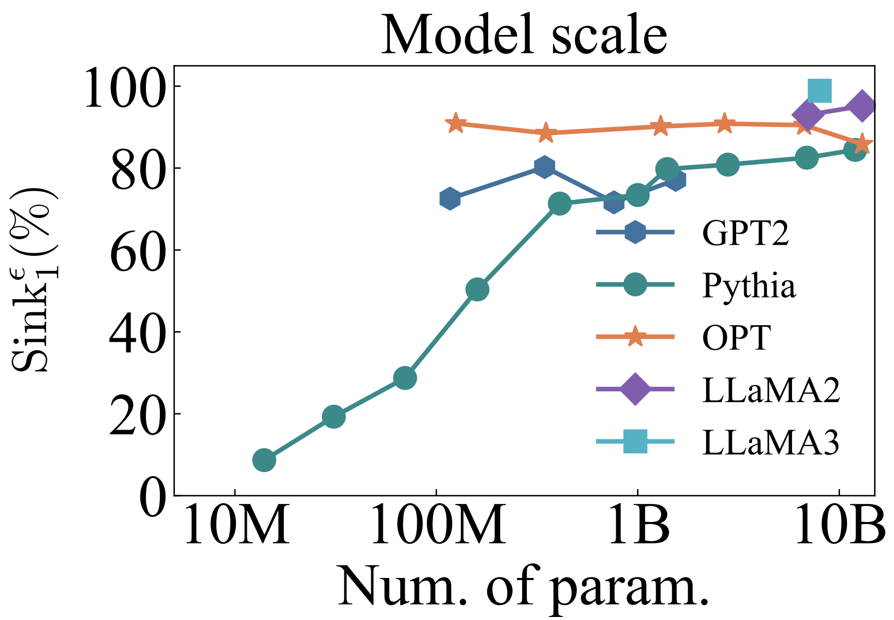
$$\mathrm{Sink}_k^{\epsilon} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{H}\sum_{h=1}^{H}\mathbb{I}(\alpha_k^{l,h} > \epsilon)$$

Attention sink metric of the whole LM

Within a head, a threshold to decide a sink

How to measure attention sink?

- Attention sink appears widespread in various LMs, even in LMs with14M params.



- Attention sink emerges in LM pre-training

LLM	$\mathrm{Sink}_1^{\epsilon}(\%)$ Base	$\mathrm{Sink}_1^{\epsilon}(\%)$ Chat
Mistral-7B	97.49	88.34
LLaMA2-7B	92.47	92.88
LLaMA2-13B	91.69	90.94
LLaMA3-8B	99.02	98.85

How to measure attention sink?

- Attention sink appears with / without BOS, even appears under random tokens
- Under all the repeat token input?

LLM	$\text{Sink}_1^{\epsilon}(\%)$ natural	random	repeat
GPT2-XL	77.00	70.29	62.28
Mistral-7B	97.49	75.21	0.00
LLaMA2-7B Base	92.47	90.13	0.00
LLaMA3-8B Base	99.02	91.23	0.00

- Models with NoPE / relative PE / ALiBi / Rotary have same hidden states while models with absolute / learnable PE do not

Impact of positional embeddings under repeated tokens

- Closed form/upper bound for NoPE / relative PE / ALiBi / Rotary

Proposition 1. *For LMs with NoPE, the attention scores for t repeated tokens are t^{-1} uniformly, i.e., there is no attention sink.*

Proof. We have that

$$\boldsymbol{A}_{ti}^{l,h} = \frac{e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle}}{\sum_{j=1}^{t} e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_j^{l,h} \rangle}} = \frac{e^{\boldsymbol{q}_t^{l,h} \boldsymbol{k}_i^{l,h\top}}}{\sum_{j=1}^{t} e^{\boldsymbol{q}_t^{l,h} \boldsymbol{k}_j^{l,h\top}}} = \frac{e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top}}}{t e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top}}} = \frac{1}{t}. \tag{18}$$

Therefore, the attention scores follow a uniform distribution over all previous tokens. $\square$

Proposition 2. *For LMs with relative PE, there is no attention sink for t repeated tokens.*

Proof. For LMs with relative PE, the dot product between each query and key is

$$\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle = \boldsymbol{q}_t^{l,h} \boldsymbol{k}_i^{l,h\top} + g_{\text{rel}}(t-i) = \boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{rel}}(t-i), \tag{19}$$

then we have the attention scores

$$\boldsymbol{A}_{t,i}^{l,h} = \frac{e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle}}{\sum_{j=1}^{t} e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_j^{l,h} \rangle}} = \frac{e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{rel}}(t-i)}}{\sum_{j=1}^{t} e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{rel}}(t-j)}} = \frac{e^{g_{\text{rel}}(t-i)}}{\sum_{j=1}^{t} e^{g_{\text{rel}}(t-j)}}. \tag{20}$$

Impact of positional embeddings under repeated tokens

- Closed form/upper bound for NoPE / relative PE / ALiBi / Rotary

Proposition 3. *For LMs with ALiBi, there is no attention sink for t repeated tokens.*

Proof. For LMs with ALiBi, similar to relative PE, the dot product between each query and key is

$$\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle = \boldsymbol{q}_t^{l,h} \boldsymbol{k}_i^{l,h\top} + g_{\text{alibi}}^h(t-i) = \boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{alibi}}^h(t-i), \tag{21}$$

then we have the attention scores

$$\boldsymbol{A}_{t,i}^{l,h} = \frac{e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle}}{\sum_{j=1}^t e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_j^{l,h} \rangle}} = \frac{e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{alibi}}^h(t-i)}}{\sum_{j=1}^t e^{\boldsymbol{q}^{l,h} \boldsymbol{k}^{l,h\top} + g_{\text{alibi}}^h(t-j)}} = \frac{e^{g_{\text{alibi}}^h(t-i)}}{\sum_{j=1}^t e^{g_{\text{alibi}}^h(t-j)}}. \tag{22}$$

Here $g_{\text{alibi}}^h(t-i)$ is monotonic decreasing function of $t-i$, so there is no attention sink on the first token. $\square$

Impact of positional embeddings under repeated tokens

- Closed form/upper bound for NoPE / relative PE / ALiBi / Rotary

Proof. For LMs with Rotary, the dot product between each query and key is

$$\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle = \boldsymbol{q}_t^{l,h} \boldsymbol{R}_{\Theta,\, i-t} \boldsymbol{k}_i^{l,h\top} \tag{23}$$

$$= \boldsymbol{q}^{l,h} \boldsymbol{R}_{\Theta,\, i-t} \boldsymbol{k}^{l,h\top} \tag{24}$$

$$= \left\| \boldsymbol{q}^{l,h} \right\| \left\| \boldsymbol{k}^{l,h} \boldsymbol{R}_{\Theta,\, t-i} \right\| \cos \left(\frac{\boldsymbol{q}^{l,h} \boldsymbol{R}_{\Theta,\, i-t} \boldsymbol{k}^{l,h\top}}{\left\| \boldsymbol{q}^{l,h} \right\| \left\| \boldsymbol{k}^{l,h} \boldsymbol{R}_{\Theta,\, t-i} \right\|} \right) \tag{25}$$

$$= \left\| \boldsymbol{q}^{l,h} \right\| \left\| \boldsymbol{k}^{l,h} \right\| \cos(\beta_{t-i}), \tag{26}$$

where β_{j-t} is the angle between the rotated query and the rotated key. Then the attention scores are

$$\boldsymbol{A}_{t,\, i}^{l,\, h} = \frac{e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_i^{l,h} \rangle}}{\sum_{j=1}^{t} e^{\langle \boldsymbol{q}_t^{l,h}, \boldsymbol{k}_j^{l,h} \rangle}} = \frac{e^{\boldsymbol{q}^{l,h} \boldsymbol{R}_{\Theta,\, j-i} \boldsymbol{k}^{l,h\top}}}{\sum_{j=1}^{t} e^{\boldsymbol{q}^{l,h} \boldsymbol{R}_{\Theta,\, j-i} \boldsymbol{k}^{l,h\top}}} = \frac{e^{\left\| \boldsymbol{q}^{l,h} \right\| \left\| \boldsymbol{k}^{l,h} \right\| \cos(\beta_{t-i})}}{\sum_{j=1}^{t} e^{\| \boldsymbol{q}^{l,h} \| \| \boldsymbol{k}^{l,h} \| \cos(\beta_{t-j})}}. \tag{27}$$

Suppose the norm of multiplication for query and key $\left\| \boldsymbol{q}^{l,h} \right\| \left\| \boldsymbol{k}^{l,h} \right\| = \xi$. Considering $-1 \leq \cos(\beta_{t-j}) \leq 1$, then we have

$$\boldsymbol{A}_{t,\, i}^{l,\, h} = \frac{e^{\xi \cos(\beta_{t-i})}}{\sum_{j=1}^{t} e^{\xi \cos(\beta_{t-j})}} = \frac{1}{1 + \frac{\sum_{j \neq i} e^{\xi \cos(\beta_{t-j})}}{e^{\xi \cos(\beta_{t-i})}}} \leq \frac{e^{2\xi}}{e^{2\xi} + (t-1)} \tag{28}$$

Then the attention scores for each token are upper-bounded and decrease to 0 as t grows. $\square$

Attributing attention sink to LM pre-training

- LM pre-training objective $$\min_{\theta} \mathbb{E}_{\boldsymbol{X} \sim p_{\text{data}}} \left[\mathcal{L} \left(p_{\theta} (\boldsymbol{X}) \right) \right]$$

- Experiments on LLaMA2-style models

Optimization

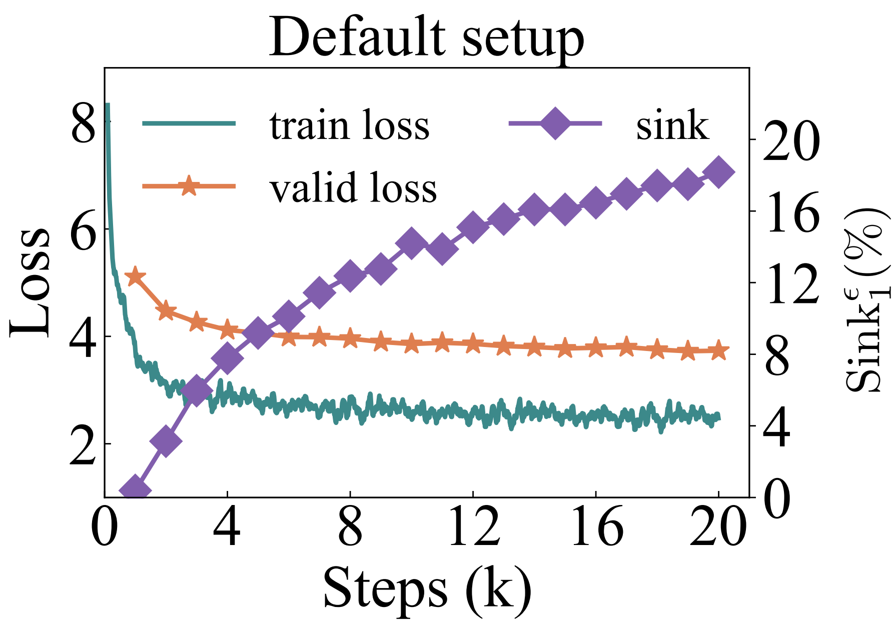
Data distribution

Loss function

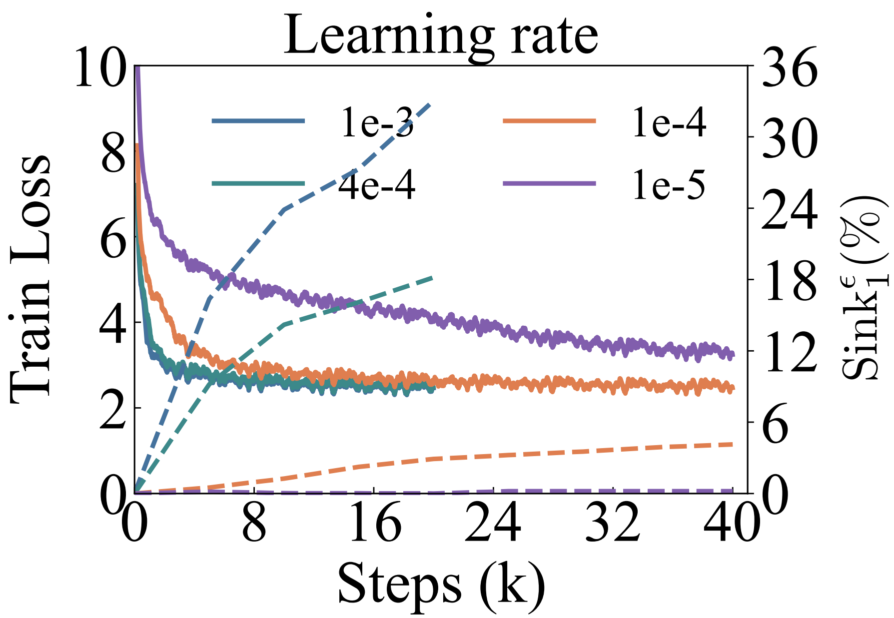
Model architecture

Effects of optimization on attention sink

- Training steps



- Learning rate



Effects of optimization on attention sink

- Small learning rates not only slow down the emergence, but also mitigate attention sink

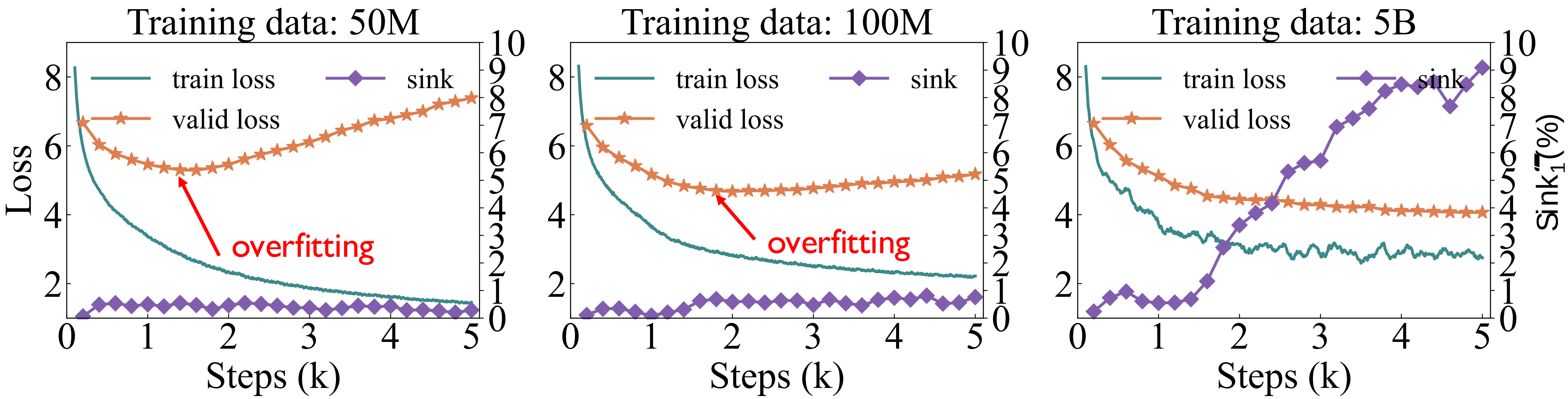
learning rate	training steps (k)	$\mathrm{Sink}_1^{\epsilon}(\%)$	valid loss
8e-4	10	23.44	3.79
8e-4	20	32.23	3.70
4e-4	20	18.18	3.73
2e-4	20	11.21	3.78
2e-4	40	16.81	3.68
1e-4	20	2.90	3.92
1e-4	80	6.29	3.67

We keep the training steps x learning rate the same

Effects of data distribution on attention sink

- Unique training data amount

Attention sink emerges after LMs are trained on sufficient unique training data, not really related to overfitting



Effects of loss function on attention sink

- Auto-regressive loss

$$\mathcal{L} = \sum_{t=2}^{C} \log p_\theta(\boldsymbol{x}_t | \boldsymbol{x}_{<t})$$

- Weight decay

L2 regularization

$$\mathcal{L} = \sum_{t=2}^{C} \log p_\theta(\boldsymbol{x}_t | \boldsymbol{x}_{<t}) + \gamma \|\theta\|_2^2$$

Larger weight decay encourages attention sink

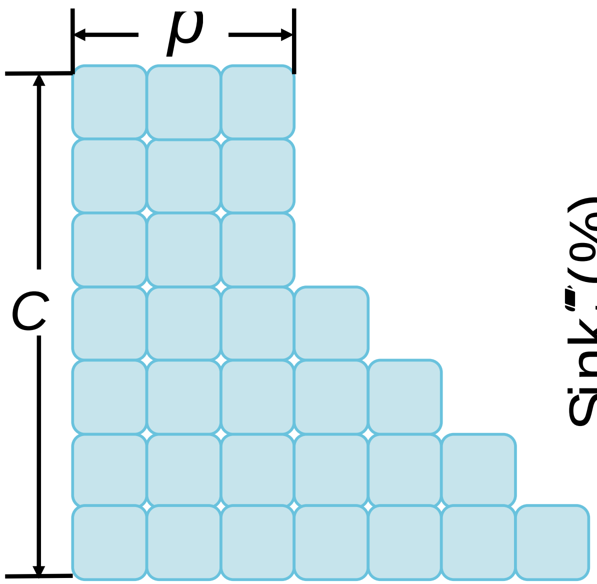
γ	0.0	0.001	0.01	0.1	0.5	1.0	2.0	5.0
$\text{Sink}_1^{\epsilon}(\%)$	15.20	15.39	15.23	18.18	41.08	37.71	6.13	0.01
valid loss	3.72	3.72	3.72	3.73	3.80	3.90	4.23	5.24

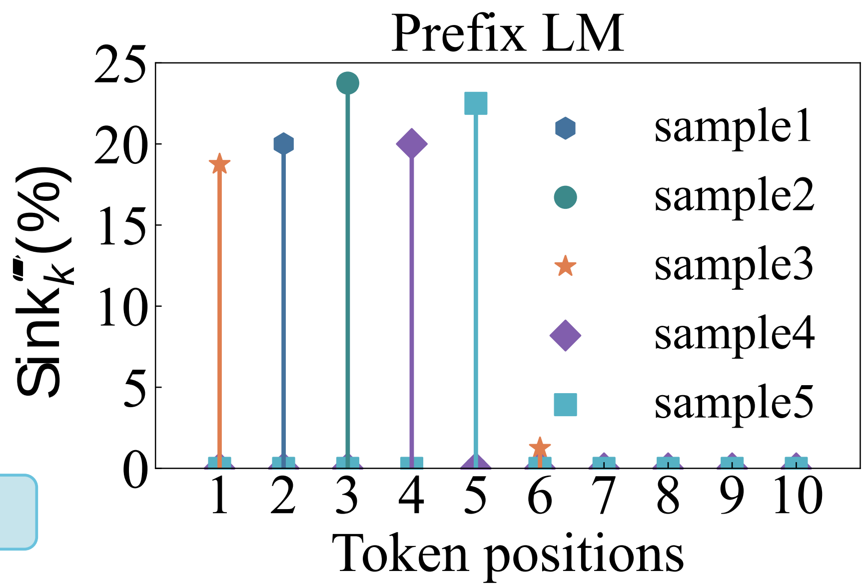
Effects of loss function on attention sink

More prefix tokens

- Prefix language modeling $\mathcal{L} = \sum_{t=p+1}^{C} \log p_\theta(\boldsymbol{x}_t | \boldsymbol{x}_{p+1:t-1}, \boldsymbol{x}_{1:p})$

Sink token shifts from the first position to other positions within the prefix





Effects of loss function on attention sink

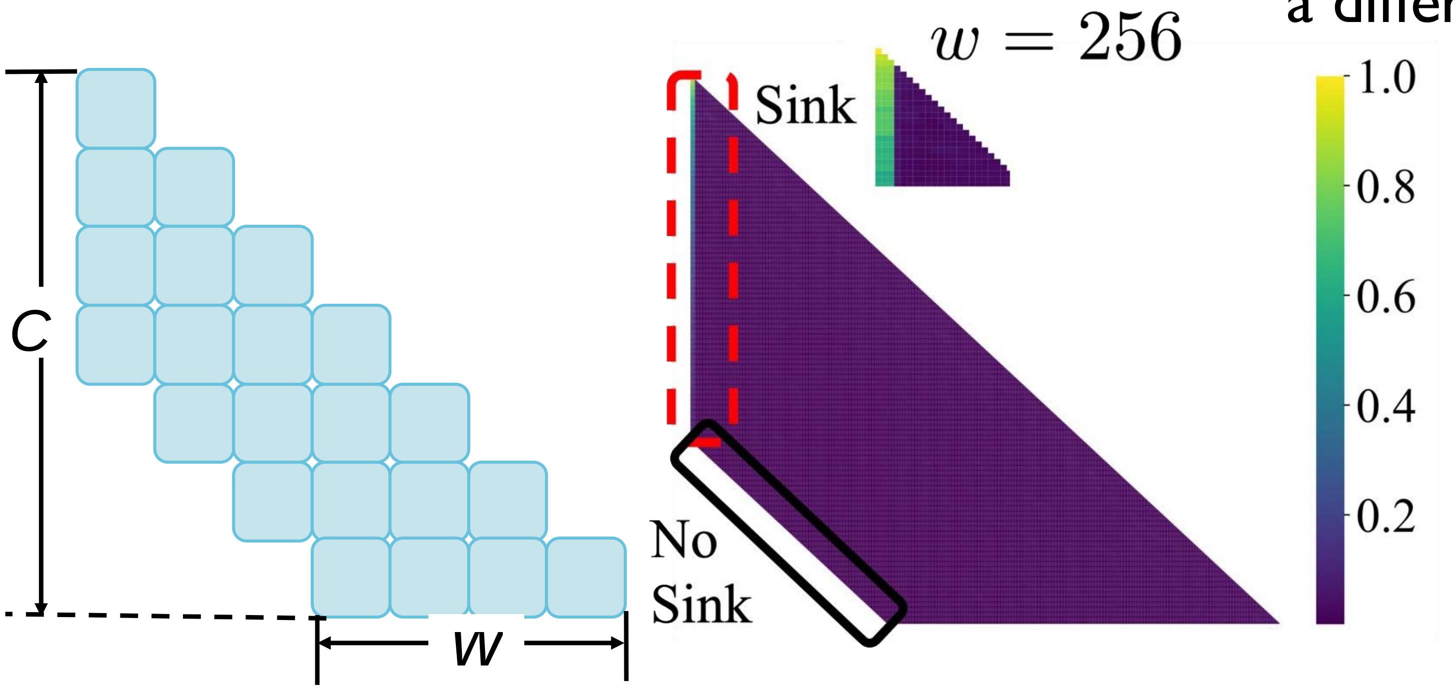
- Shifted window attention

$$\mathcal{L} = \sum_{t=2}^{C} \log p_\theta(\boldsymbol{x}_t | \boldsymbol{x}_{t-w:t-1})$$

Attention sink appears on the absolute, not the relative first token

Key of the sink token is trained to be distributed in a different manifold

Small window size mitigates attention sink



Effects of model architecture on attention sink

The following designs do not affect the emergence of attention sink

- Positional embeddings: including no positional embedding
- Pre-norm and post-norm transformer block structure
- Feed forward networks (FFNs) with different activation functions
- Number of attention heads, how to combine multiple heads

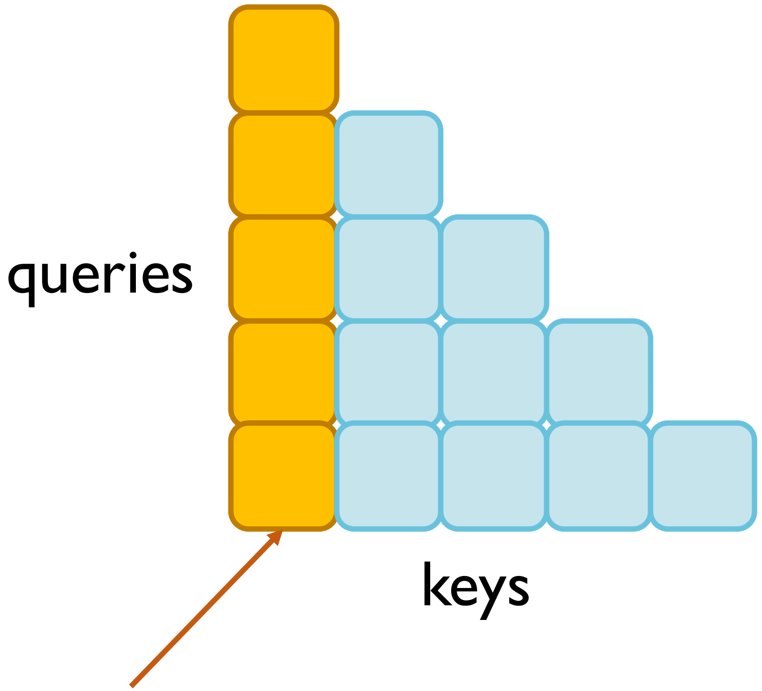
Effects of model architecture on attention sink

Standard softmax attention in h-th head l-th block

$$\text{Softmax}\left(\frac{1}{\sqrt{d_h}}\boldsymbol{Q}^{l,h}{\boldsymbol{K}^{l,h}}^{\top}+\boldsymbol{M}\right)\boldsymbol{V}^{l,h}$$

queries keys values

casual mask



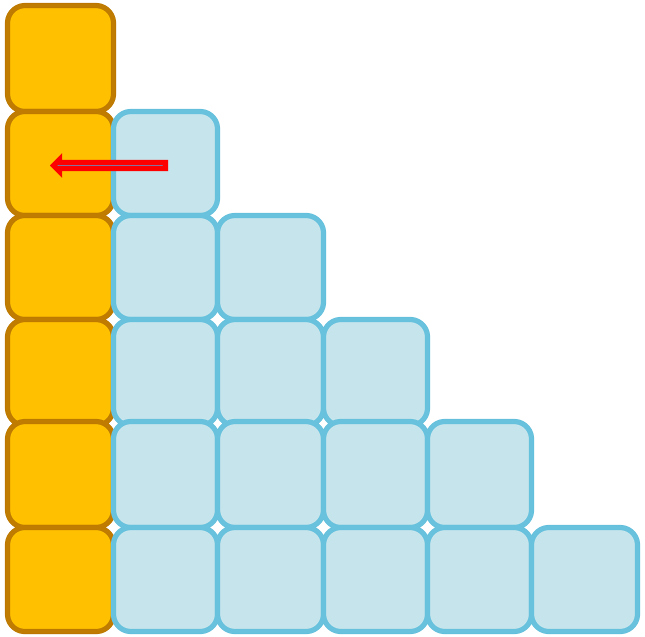
Effects of model architecture on attention sink

Softmax attention with a learnable sink token (Xiao et al. 2024)

$$\mathrm{Softmax}\left(\frac{1}{\sqrt{d_h}}\begin{bmatrix}\boldsymbol{q}^{*l,h}\\ \boldsymbol{Q}^{l,h}\end{bmatrix}\begin{bmatrix}{\boldsymbol{k}^{*l,h}}^\top & {\boldsymbol{K}^{l,h}}^\top\end{bmatrix}+\boldsymbol{M}\right)\begin{bmatrix}\boldsymbol{v}^{*l,h}\\ \boldsymbol{V}^{l,h}\end{bmatrix}$$

QKV for sink token

queries



keys

$\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \boldsymbol{x}_3 \quad \cdots \quad \boldsymbol{x}_T$

$\boldsymbol{x}^* \quad \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \boldsymbol{x}_3 \quad \cdots \quad \boldsymbol{x}_T$

Sink on the learnable sink token

Learnable sink token

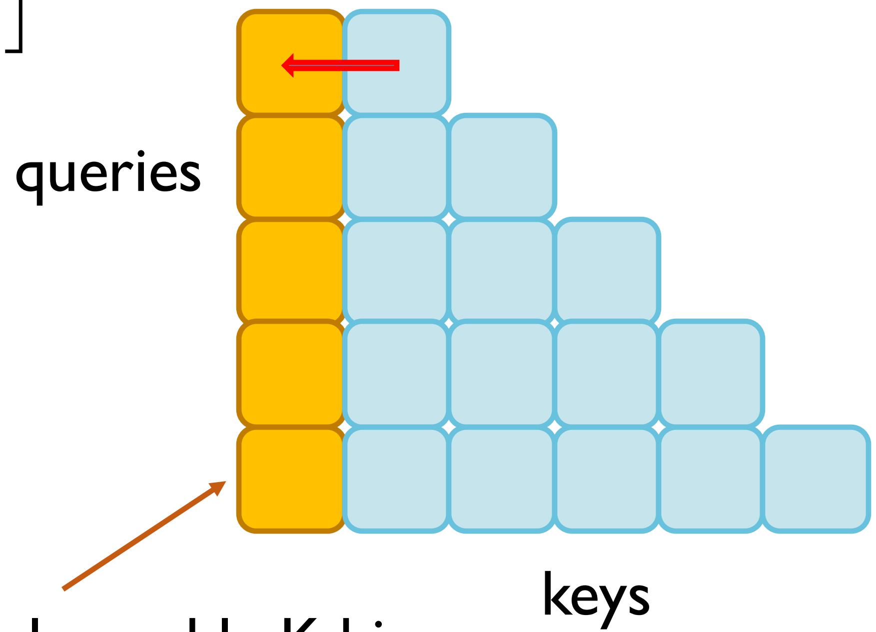
Xiao et al. Efficient Streaming Language Models with Attention Sinks. ICLR 2024

Effects of model architecture on attention sink

Softmax attention with learnable KV biases (Sun et al. 2024)

$$\mathrm{Softmax}\left(\frac{1}{\sqrt{d_h}}\boldsymbol{Q}^{l,h}\begin{bmatrix}{\boldsymbol{k}^{*l,h}}^\top & {\boldsymbol{K}^{l,h}}^\top\end{bmatrix} + \boldsymbol{M}\right)\begin{bmatrix}\boldsymbol{v}^{*l,h} \\ \boldsymbol{V}^{l,h}\end{bmatrix}$$

Learnable KV biases



Sink on the learnable K bias

Sun et al. Massive activations in large language models. COLM 2024

Effects of model architecture on attention sink

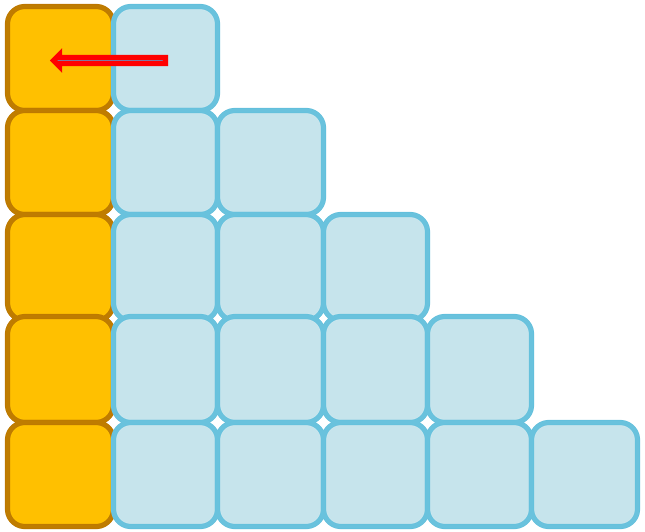
Softmax attention with learnable K biases

$$\mathrm{Softmax}\left(\frac{1}{\sqrt{d_h}}\boldsymbol{Q}^{l,h}\begin{bmatrix}\boldsymbol{k}^{*l,h^\top} & \boldsymbol{K}^{l,h^\top}\end{bmatrix}+\boldsymbol{M}\right)\begin{bmatrix}\mathbf{0}\\ \boldsymbol{V}^{l,h}\end{bmatrix}$$

V biases are all zeros

Learnable K biases

queries



keys

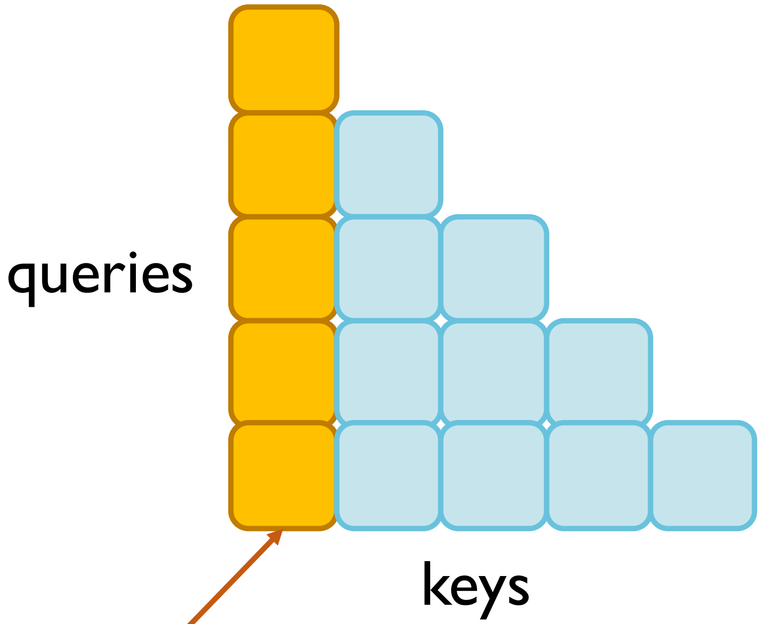
Sink on the learnable K bias

Effects of model architecture on attention sink

Softmax attention with learnable V biases (control group)

$$\mathrm{Softmax}\left(\frac{1}{\sqrt{d_h}}\boldsymbol{Q}^{l,h}{\boldsymbol{K}^{l,h}}^{\top}+\boldsymbol{M}\right)\boldsymbol{V}^{l,h}+\boldsymbol{v}^{*l,h}$$

Learnable V biases



Sink on the first token,
no effects

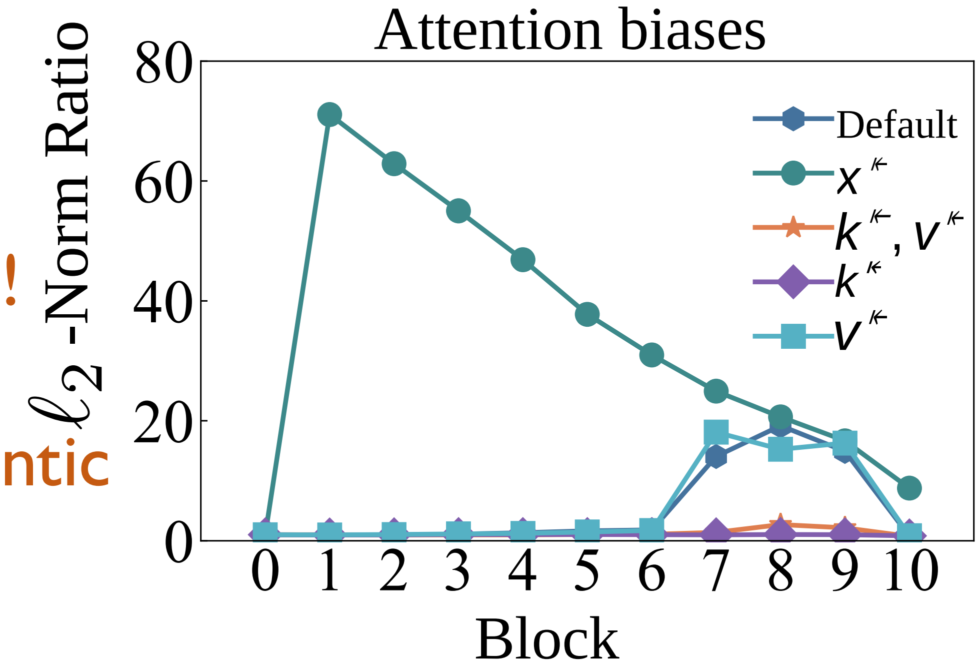
Effects of model architecture on attention sink

Hidden states' L2-norm ratios between the first token and other tokens

- LM with K biases has no massive activations!

- Large attention score $\neq$ important in semantic

- Sink token saves extra attention, adjusts the dependence among other tokens



Why need such a mechanism?

Is it because attention score added up to one?

Effects of model architecture on attention sink

Attention output

$$\mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_j)) = \exp(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d_h}})$$

softmax

$$\boldsymbol{v}_i^\dagger = \sum_{j=1}^{i} \frac{\mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_j))}{\sum_{j'=1}^{i} \mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_{j'}))} \boldsymbol{v}_j$$

$$\boldsymbol{Z}_i = \sum_{j'=1}^{i} \mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_j))$$

normalization term

Perhaps normalization matters, as it forces the attention scores sum to one?

Effects of model architecture on attention sink

- Scale the normalization term

$$\boldsymbol{Z}_i \rightarrow \boldsymbol{Z}_i/\alpha$$

- Power of attention scores sum up to one

$$\boldsymbol{v}_i^{\dagger} = \frac{\sum_{j=1}^{i} \mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_j))\boldsymbol{v}_j}{\left(\sum_{j'=1}^{i} \mathrm{sim}(\varphi(\boldsymbol{q}_i), \varphi(\boldsymbol{k}_{j'}))^p\right)^{\frac{1}{p}}}$$

$$\boldsymbol{v}_i^{\dagger} = \sum_{j=1}^{i} \left(\frac{\exp(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\top}}{\sqrt{d_h}/p})}{\sum_{j'=1}^{i} \exp(\frac{\boldsymbol{q}_i \boldsymbol{k}_{j'}^{\top}}{\sqrt{d_h}/p})} \right)^{\frac{1}{p}} \boldsymbol{v}_j$$

softmax

- May mitigate attention sink, but not prevent the emergence

Effects of model architecture on attention sink

- Relax tokens' inner dependence by removing normalization

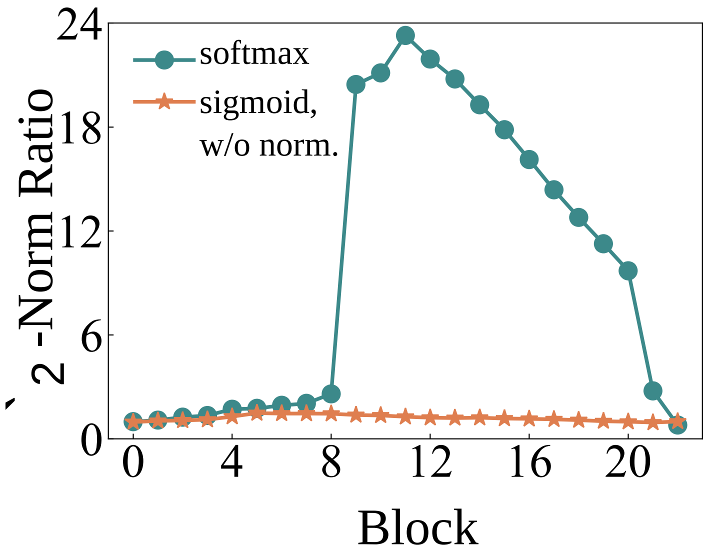
Sigmoid attention:

$$\boldsymbol{v}_i^{\dagger} = \sum_{j=1}^{i} \mathrm{sigmoid}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\top}}{\sqrt{d_h}}\right)\boldsymbol{v}_j$$

ELU plus one attention:

$$\boldsymbol{v}_i^{\dagger} = \sum_{j=1}^{i} \left(\mathrm{elu}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\top}}{\sqrt{d_h}}\right) + 1\right)\boldsymbol{v}_j$$

Scale up to 1B LLMs



No normalization -> No attention sink, no massive activations!
Added back normalization -> Attention sink, massive activations!

Effects of model architecture on attention sink

- Relax tokens' inner dependence by allowing negative attention scores

Linear attention, with a mlp kernel

$$\boldsymbol{v}_i^\dagger = \sum_{j=1}^{i} \frac{\mathrm{mlp}(\boldsymbol{q}_i)\mathrm{mlp}(\boldsymbol{k}_j)^\top}{\sqrt{d_h}} \boldsymbol{v}_j$$

-> No attention sink, no massive activations

Add a normalization

$$\boldsymbol{Z}_i = \max\left(\left| \sum_{j'=1}^{i} \frac{\mathrm{mlp}(\boldsymbol{q}_i)\mathrm{mlp}(\boldsymbol{k}_{j'})^\top}{\sqrt{d_h}} \right|, 1 \right)$$

-> No attention sink, no massive activations

Takeaway

- Attention sink is a widespread phenomena across models and input
- Attention sink emerges during the LM pre-training
- Attention sink acts as key biases, storing extra attention and non-informative
- Softmax plays an important role in the emergence of attention sink

Please check our paper to see more interesting results!